O

Online Controlled Experiments and A/B Tests

Ron Kohavi[1] and Roger Longbotham[2]
[1]Kohavi, Los Altos, CA, USA
[2]Process Performance Management, Meridian, ID, USA

Motivation and Background

Many good resources are available with motivation and explanations about online controlled experiments (Kohavi et al. 2009a, 2020; Thomke 2020; Luca and Bazerman 2020; Georgiev 2018, 2019; Kohavi and Thomke 2017; Siroker and Koomen 2013; Goward 2012; Schrage 2014; King et al. 2017; McFarland 2012; Manzi 2012; Tang et al. 2010). For organizations running online controlled experiments at scale, Gupta et al. (2019) provide an advanced set of challenges.

We provide a motivating visual example of a controlled experiment that ran at Microsoft's Bing. The team wanted to add a feature allowing advertisers to provide links to the target site. The rationale is that this will improve ads quality by giving users more information about what the advertiser's site provides and allow users to directly navigate to the sub-category matching their intent. Visuals of the existing ads layout (Control) and the new ads layout (Treatment) with site links added are shown in Fig. 1.

In a controlled experiment, users are randomly split between the variants (e.g., the two different ads layouts) in a persistent manner (a user receives the same experience in multiple visits). Their interactions with the site are instrumented and key metrics computed. In this experiment, the Overall Evaluation Criterion (OEC) was simple: increasing average revenue per user to Bing without degrading key user engagement metrics. Results showed that the newly added site links increased revenue, but also degraded user metrics and Page-Load-Time, likely because of increased vertical space usage. Even offsetting the space by lowering the average number of mainline ads shown per query, this feature improved revenue by tens of millions of dollars per year with neutral user impact, resulting in extremely high ROI (Return-On-Investment).

Why Experiment? Correlations, Causality, and the Hierarchy of Evidence

This section is adapted from Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing (Kohavi et al. 2020).

You introduce a new feature to your software product and notice that the user churn rate (those ending their subscription) of users utilizing

D. Phung et al. (eds.), *Encyclopedia of Machine Learning and Data Science*,
https://doi.org/10.1007/978-1-4899-7502-7_891-2

Online Controlled Experiments and A/B Tests, Fig. 1 Ads with site link experiment. Treatment (bottom) has site links. The difference might not be obvious at first, but it is worth tens of millions of dollars per year to Bing

the feature decreases dramatically. You might be tempted to claim causality; the feature is reducing churn. You might claim that if we make the feature more discoverable and used more often, subscriptions will soar. Wrong! Given the data, no conclusion can be drawn about whether the feature reduces or increases user churn, and both are possible.

An example demonstrating this fallacy comes from Microsoft Office 365. Users of Office 365 that see error messages and experience crashes have lower churn rates, but that does not mean that Office 365 should show more error messages or that Microsoft should lower code quality, causing more crashes. It turns out that all three events are caused by a single factor: usage. Heavy users of the product see more error messages, experience more crashes, and have lower churn rates. Correlation does not imply causality and overly relying on these observations leads to faulty decisions.

Guyatt et al. (1995) introduced the hierarchy of evidence as a way to grade recommendations in medical literature, which Greenhalgh expanded on in her discussions on practicing evidence-based medicine (2014). Randomized controlled experiments are the gold standard for establishing causality. Systematic reviews, that is, meta-analysis, of controlled experiments provides more evidence and generalizability. Below that, the trust-level reduces dramatically: you can have controlled experiments that are not randomized, observational studies (cohort and case control), and case studies, anecdotes, and personal (often expert) opinions.

Online controlled experiments are:

- The best scientific way to establish causality with high probability.
- Able to detect small changes that are harder to detect with other techniques, due to changes over time or correlation with other factors (sensitivity).
- Able to detect unexpected changes. Often underappreciated, but many experiments uncover surprising impacts on other metrics, such as performance degradation, increased crashes/errors, or cannibalizing clicks from other features.

Online controlled experiments provide an unparalleled ability to electronically collect reliable data at scale, randomize well, and avoid or detect pitfalls. We recommend using other, less trustworthy methods, including observational studies, when online controlled experiments are not possible.

Key Tenets for Online Experimentation

Running online controlled experiments is not applicable for every organization. We begin with

key tenets, or other metrics, such as performance assumptions, an organization needs to adopt (Kohavi et al. 2013, 2020).

Tenet 1: The Organization Wants to Make Data-Driven Decisions and Has Formalized the Overall Evaluation Criterion (OEC)

You will rarely hear someone at the head of an organization say that they don't want to be data-driven, but measuring the incremental benefit to users from new features has costs, and objective measurements typically show that progress is not as rosy as initially envisioned. In any organization, there are many important metrics reflecting revenue, cost, customer satisfaction, loyalty, etc. and very frequently an experiment will improve one but hurt another of these metrics. Having a single metric, which we call the Overall Evaluation Criterion, or OEC, that is at a higher level than these and incorporates the trade-off among them is essential for organizational decision-making.

An OEC has to be defined and it should be measurable over relatively short durations (e.g., 2 weeks). The hard part is finding metrics that are measurable in the short-term that are predictive of long-term goals. For example, "Profit" is not a good OEC, as short-term theatrics (e.g., raising prices) can increase short-term profit, but hurt it in the long run. As shown in Trustworthy Online Controlled Experiments: Five Puzzling Outcomes Explained (Kohavi et al. 2012), market share can be a long-term goal, but it is a terrible short-term criterion: making a search engine worse forces people to issue more queries to find an answer, but, like hiking prices, users will find better alternatives long-term. Sessions per user, or repeat visits, is a much better OEC for a search engine. Thinking of the drivers of lifetime value can lead to a strategically powerful OEC (Kohavi et al. 2009a). We cannot overemphasize the importance of coming up with a good OEC that the organization can align behind. There are two chapters on Organizational Metrics and Metrics for Experimentation and the Overall Evaluation Criterion (OEC) in Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing (Kohavi et al. 2020).

Tenet 2: Controlled Experiments Can Be Run, and Their Results Are Trustworthy

Not every decision can be made with the scientific rigor of a controlled experiment. For example, you cannot run a controlled experiment on the possible acquisition of one company by another. Hardware devices may have long lead times for manufacturing and modifications are hard, so controlled experiments with actual users are hard to run on a new phone or tablet. For customer-facing web sites and services, changes are easy to make through software, and running controlled experiments is relatively easy.

With online controlled experiment, the typical experimental unit is the user or visitor of the website, app, etc. There must be enough users to detect effects due to changes. We recommend the number of users available for a test be at least in the thousands. Large sites may have hundreds of thousands or millions in a test and are able to detect small changes. However, even small sites can run A/B tests to detect moderate or large changes.

Assuming one can run controlled experiments, it is important to ensure their trustworthiness. When running online experiments, getting numbers is easy; getting numbers one can trust is hard, and we have had our share of pitfalls and puzzling results (Kohavi et al. 2010, 2012; Kohavi and Longbotham 2010; Crook et al. 2009; Deng et al. 2017; Dmitriev et al. 2016).

Tenet 3: We Are Poor at Assessing the Value of Ideas

Features are built because teams believe they are useful, yet in many domains, most ideas fail to improve key metrics. Only one third of the ideas tested on the Experimentation Platform at Microsoft improved the metric(s) they were designed to improve (Kohavi et al. 2009b). Success is even harder to find in well-optimized domains like Bing. Jim Manzi (2012) wrote that at Google, only "about 10 percent of these [controlled experiments, were] leading to business changes." Avinash Kaushik wrote in his Experimentation and Testing primer (Kaushik 2006) that "80% of the time you/we are wrong about what a customer wants." Mike Moran (2007, 240)

wrote that Netflix considers 90% of what they try to be wrong. Regis Hadiaris from Quicken Loans wrote that "in the five years I've been running tests, I'm only about as correct in guessing the results as a major league baseball player is in hitting the ball. That's right - I've been doing this for 5 years, and I can only "guess" the outcome of a test about 33% of the time!" (Moran 2008). Dan McKinley at Etsy wrote (McKinley 2013) "nearly everything fails" and "it's been humbling to realize how rare it is for them [features] to succeed on the first attempt. I strongly suspect that this experience is universal, but it is not universally recognized or acknowledged." Finally, Colin McFarland wrote in the book Experiment! (McFarland 2012, 20) "No matter how much you think it's a no-brainer, how much research you've done, or how many competitors are doing it, sometimes, more often than you might think, experiment ideas simply fail."

Not every domain has such poor statistics, but most who have run controlled experiments in customer-facing web sites and applications have experienced this humbling reality: we are poor at assessing the value of ideas, and that is the greatest motivation for getting an objective assessment of features using controlled experiments.

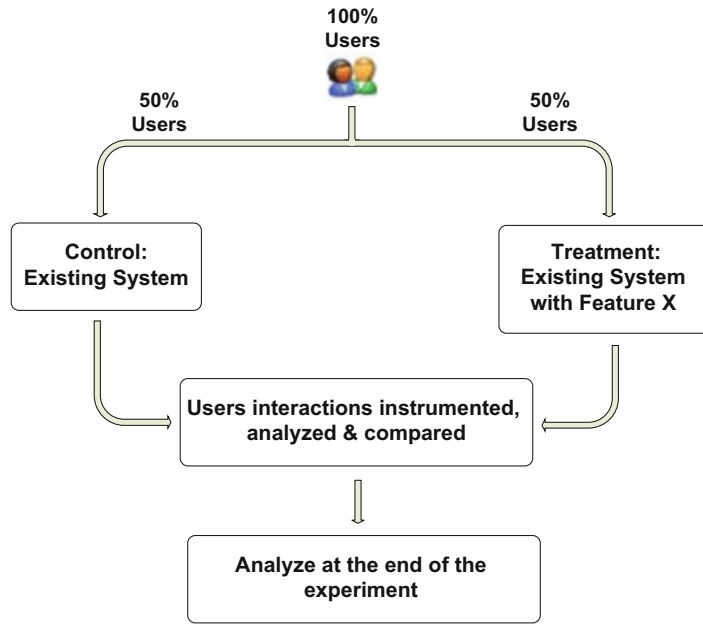
Structure of an Experimentation System

Elements of an Experimentation System

The simplest experimental setup is to evaluate a factor with two levels, a Control (version A) and a Treatment (version B). The control is normally the default version and the treatment is the change that is tested. Such a setup is commonly called an A/B test. It is commonly extended by having several levels, often referred to as A/B/n split tests. An experiment with multiple factors is referred to as Multivariable (or Multivariate).

Figure 2 shows the high-level structure of an A/B experiment. In practice, one can assign any percentages to the Treatment and Control but 50% provides the experiment the maximum statistical power, and we recommend maximally powering the experiments after a ramp-up period at smaller percentages to check for egregious errors (Kohavi et al. 2020, Chapter 15: Ramping Experiment Exposure).

In a general sense, the analysis will test if the statistical distribution of the Treatment is different from that of the Control. In practice, the most common test is whether the two means are equal or not. For this case, the effect of version *B*



Online Controlled Experiments and A/B Tests, Fig. 2 High-level structure of an online experiment

(or treatment effect) is defined to be

$$E(B) = \overline{X}_B - \overline{X}_A \quad (1)$$

Where X is a metric of interest and $\overline{X}_B$ is the mean for variant B. However, for interpretability, the percent change is normally reported with a suitable (e.g., 95%) confidence interval. See, for example Kohavi et al. (2009a). In some cases, a comparison of the variants using a statistic other than the mean may be needed. An example of when the quantiles of the distribution of the metrics would normally be preferred to the mean would be for performance metrics. However, as noted in Dmitriev et al. (2017), comparing the quantiles of variants often has less power than comparison of the means and is computationally more expensive.

Control of extraneous factors and randomization are two essential elements of any experimentation system. Any factor that may affect an online metric is either a test factor (one you intentionally vary to determine its effect) or a non-test factor. Non-test factors could either be held fixed, blocked, or randomized. Holding a factor fixed can impact external validity and is thus not recommended. For example, if weekend days are known to be different from weekdays, you could run the experiment only on weekdays (or weekends) but it would be better to have complete weeks in the experiment for better external validity. Blocking (e.g., pairing) can reduce the variance relative to randomization, and is recommended when experimentation units in each block are more homogenous than between blocks. For example, if the randomization unit is a user page view, then blocking by weekend/weekday can reduce the variance of the effect size, leading to higher sensitivity. Time is a critical non-test factor, and because many external factors vary with time, it is important to randomize over time by running the Control and Treatment(s) concurrently with a fixed percentage to each throughout the experiment. (If the relative percentage changes you will be subject to Simpson's paradox (Malinas and Bigelow 2009; Kohavi and Longbotham 2010)). Controlling a non-test factor assures it will have equal influence on the Control and Treatment, hence not affecting the estimate of the treatment effect.

Experimentation Architecture Alternatives

Controlled experiments on the web: survey and practical guide (Kohavi et al. 2009a) provides a review of architecture alternatives. The main three components of an experimentation capability involve the randomization algorithm, the assignment method (i.e., how the randomly assigned experimental units are given the variants) and the data path (which captures raw observation data and processes it). Tang et al. (2010) give a detailed view of the infrastructure for experiments as carried out by Google.

To validate an experimentation system, we recommend that A/A tests be run regularly to test that the experimental setup and randomization mechanism is working properly (Kohavi et al. 2009a, 2020, Chapter 19: The A/A Test). An A/A test, sometimes called a Null Test (Peterson 2004), exercises the experimentation system, assigning users to one of two groups, but exposes them to exactly the same experience. An A/A test can be used to (i) collect data and assess its variability for power calculations, and (ii) test the experimentation system (the Null hypothesis should be rejected about 5% of the time when a 95% confidence level is used).

Planning Experiments

Several aspects of planning an experiment are important: estimating adequate sample size, gathering the right metrics, tracking the right users, randomization unit.

Sample size. Sample size is determined by the percent of users admitted into the experiment variants (control and treatments) and how long the experiment runs. As an experiment runs longer, more visitors are admitted into the variants, so sample sizes increase. Experimenters can choose the relative percent of visitors that are in the Control and Treatment, which affects how long you will need to run the experiment. Several authors (Deng et al. 2013; Kohavi et al. 2009a) have addressed the issue of sample size

and length of experiment in order to achieve adequate statistical power for an experiment, where statistical power of an experiment is the probability of detecting a given effect when it exists. For example, one common approach to sizing an experiment (% of population in each variant and length of time) is to have an 80% chance of achieving statistical significance if the treatment is different from the control by X%. In addition to planning an experiment for adequate power, a best practice is to run the experiment for at least 1 week (to capture a full weekly cycle) and then multiple weeks beyond that. When "novelty" or "primacy" effects are suspected (i.e., the initial effect of the Treatment is not the same as the long-term effect), the experiment should be run long enough to estimate the asymptotic effect of the Treatment.

Note that statistical power varies from metric to metric. An experiment with 100,000 users running for 2 weeks may have sufficient power to detect a 1% change in one metric but only able to detect a 3% change in a different metric. We have found that some metrics, such as sessions per user, will require more users than other metrics to achieve the same power to detect a certain percent difference. Generally, we expect the power of an experiment for a given metric to increase as the experiment continues. However, this also varies from metric to metric. Two metrics can have different power profiles over time. For example, click-through rate (CTR), defined as number of clicks divided by the number of pageviews, has increasing power as the experiment runs longer, but sessions per user does not (Kohavi et al. 2012).

Observations, Metrics, and the OEC. Gathering observations (i.e., logging events) so that the right metrics can be computed is critical to successful experimentation. Whenever possible and economically feasible, one should gather as many observations as possible that relate to answering potential questions of interest, whether user related or performance related (e.g., latency, utilization, crashes). We recommend computing many metrics from the observations (e.g., hundreds) because they can give rise to surprising insights, although care must be taken to correctly understand and control for the false positive rate. While having many metrics is great for insights, decisions should be made using the Overall Evaluation Criterion (OEC). See Tenet 1 earlier for a description of the OEC.

Triggering. Some treatments may be relevant to all users who come to a website. However, for many experiments, the difference introduced is relevant for a subset of visitors (e.g., a change to the checkout process, which only 10% of visitors start). In these cases, it is best to include only those visitors who would have experienced a difference in one of the variants (this commonly requires counter-factual triggering for the control). Some architectures (Kohavi et al. 2009a) trigger users into an experiment either explicitly or using lazy (or late-bound) assignment. In either case, the key is to analyze only the subset of the population that was potentially impacted. Triggering reduces the variability in the estimate of treatment effect, leading to more precise estimates. Because the diluted effect is often of interest, the effect can then be diluted (Deng and Hu 2015).

Randomization Unit. Most experiments use the visitor as the randomization unit for several reasons. First, for many changes being tested it is important to give the user a consistent online experience. Second, most experimenters evaluate metrics at the user level, such as sessions per user and clicks per user. Ideally, the randomization by the experimenter is by a true user, but in many unauthenticated sites, a cookie stored by the user's browser is used, so in effect, the randomization unit is the cookie. In this case, the same user will appear to be different users if she comes to the site using a different browser, different device, or having deleted her cookie during the experiment. The next section will discuss how the choice of randomization unit affects how the analysis of different metrics should be carried out. The randomization unit can also affect the power of the test for some metrics. For example, Deng et al. (2011) showed that the variance of page level metrics can be greatly reduced if randomization is done at the page level, but user metrics cannot be computed in such cases. In social-network settings, spillover effects violate

the standard no-interference assumption, requiring unique approaches, such as clustering (Xu et al. 2015; Ugander et al. 2013; Katzir et al. 2012; Eckles et al. 2017).

Analysis of Experiments

If an experiment is carried out correctly, the analysis should be a straight-forward application of well-known statistical methods. Of course, this is much preferred than trying to recover from a poor experimental design or implementation. Dmitriev et al. (2017) provide common pitfalls of analysis and interpretation of A/B tests.

Confidence Intervals. Most reporting systems will display the treatment effect (actual and percent change) along with suitable confidence intervals. For reasonably large sample sizes, generally considered to be thousands of users in each variant the means may be considered to have normal distributions (See Kohavi et al. (2014) for detailed guidance) making the formation of confidence intervals routine. However, care must be taken to use the Fieller theorem formula (Fieller 1954) for percent effect since there is a random quantity in the denominator.

Decision-making. A common approach to deciding if the Treatment is better than the Control is the usual hypothesis-testing procedure, assuming the Normal distribution if the sample size is sufficient (Kohavi et al. 2009a). Alternatives to this when normality cannot be assumed are transformations of the data (Bickel and Doksum 1981) and nonparametric or resampling/permutation methods to determine how unusual the observed sample is under the null hypothesis (Good 2005). When conducting a test of whether the Treatment had an effect or not (e.g., a test of whether the Treatment and Control means are equal) a p value of the statistical test is often produced as evidence. More precisely, the p value is the probability to obtain an effect equal to or more extreme than the one observed, presuming the null hypothesis of no effect is true (Biau et al. 2010).

Another alternative is to use Bayes' theorem to calculate the posterior odds that the Treatment had a positive impact versus the odds it had no impact (Stone 2013).

Analysis Units. Metrics may be defined with different analysis units, such as user, session or other appropriate basis. For example, an ecommerce site may be interested in metrics such as revenue per user, revenue per session or revenue per purchaser. Straightforward statistical methods (e.g., the usual t-test and variants) apply to any metric that has user as its analysis unit if users are the unit of randomization since users may be considered independent. However, if the analysis unit is not the same as the randomization unit, the analysis units may not be considered independent and other methods need to be used to calculate standard deviation or to compare Treatment to Control. Bootstrapping (Bradley and Tibshirani 1993) and the delta method (Casella and Berger 2001; Deng et al. 2018) are two commonly used methods when the analysis unit is not the same as the randomization unit.

Increasing Experiment Sensitivity. In statistical terms, increasing sensitivity is referred to as increasing the power of the experiment. In A/B tests, the sensitivity of the experiment is often stated as the ability to detect a certain (real) percent change in a metric with a given probability. This is most helpful in planning the experiment. Various methods for increasing the sensitivity, or power, of an experiment are given below:

1. Increasing the number of randomization units (e.g., users) in the experiment (i.e., increasing sample size). Sometimes, randomizing by users every day, or by session, or even by pageview can increase sensitivity at the expense of inconsistency to users and the inability to use user-level metrics (e.g., for display ads, randomizing by page and optimizing for clicks per page suffices; for a the overall purchasing experience, revenue per user and average user time to complete checkout require randomization by user). For convenience, we assume the randomization unit is users below. Running an experiment longer, and admitting more users, will increase sensitivity for most metrics.
2. If there are two variants (i.e., Treatment and Control), having each with half the users in the

O

experiment gives the most power for a given number of users. If there are more than two variants, giving more users to Control gives more sensitivity for each test of Treatment versus Control, but may introduce problems of unequal variants due to caching (Kohavi et al. 2020, Chapter 19: The A/A Test).

3. Choosing a more sensitive metric. For example, for eCommerce sites, Revenue per User will typically have much less power (to see a certain percent change) than the Boolean, Conversion Rate (percent of visitors who purchased).

   Breaking up a metric such as Revenue per User into its two parts, (a) Conversion Rate and (b) Revenue per Purchaser can improve sensitivity. Typically, both will have more power than the original metric and give the complete picture for the metric you are interested in, Revenue per User.
4. Transforming the metric of interest. Some transformations that will, in general, improve the sensitivity of a non-Boolean metric are:
   (a) Outlier removal. Choose a value (or percentile) and remove all values above the value. For example, "users" that click more than 1000 times an hour are likely to be automated bots.
   (b) Capping. Choose a value and replace larger values with the capped value.

   The "user" may be a rare commercial entity buying a large number of items, which can skew the average of one variant by chance, not because of the change.
   (c) Log transformation. This is helpful for metrics that only take positive values and are skewed right, such as duration.
5. Analyzing triggered users. Users that could not have been impacted by the change could not have been impacted. Adding them to the analysis adds noise.
6. Using pre-experiment data to define covariates and stratification. These methods, such as CUPED (Deng et al. 2013), often reduce variation of the tested metric. Xie and Aurisset (2016) showed how these methods reduce variation of Netflix metrics.

Experimentation Trustworthiness: Common Pitfalls

This section is adapted from Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing (Kohavi et al. 2020, Chapter 3). Several "intuition busters" that are common pitfalls in the industry are shown in Kohavi, Deng, and Vermeer (2022).

When we see a surprisingly positive result, such as a significant improvement to a key metric, the inclination is to build a story around it, share it, and celebrate. When the result is surprisingly negative, the inclination is to find some limitation of the study or a minor flaw and dismiss it. Experience tells us that many extreme results are more likely to be the result of an error in instrumentation (e.g., logging), loss of data (or duplication of data), or a computational error.

Here are some tests to increase trust in the experiment results

1. Are users exposed only to one variant for an experiment? If many users are exposed to both control and a treatment, there is contamination, which is a red flag.
2. Does the experiment have sufficient statistical power? An evaluation of 115 A/B tests at GoodUI.org suggests that most were underpowered (Georgiev 2018).
3. Is there a "sample ratio mismatch" or SRM? If the ratio of users (or any randomization unit) between the variants is not close to the designed ratio, the experiment suffers from a Sample Ratio Mismatch (SRM). For example, if the experiment design is for a ratio of one-to-one (equally sized Control and Treatment), then deviations in the actual ratio of users in an experiment likely indicate a problem (Fabijan et al. 2019). One should always include a statistical comparison of the planned versus the actual percentages in the variants and give a warning if the actual is too far from planned. The experimentation system should generate a strong warning and hide any scorecards and reports if the p-value for the ratio is very low (e.g., <0.001).

4. Are triggered analyses done, if the experiment impacted a small subset of the population?
5. Are corrections done for multiple hypothesis tests? If there are multiple treatments, hundreds of metrics, multiple iterations of an experiment, small p-values are likely to occur by chance. It is important to interpret experiments with a clear understanding of p-values (Vickers 2009). Misinterpretation of p-values is quite common and can often lead to incorrect decision-making, as highlighted in A Dirty Dozen: Twelve P-Value Misconceptions (Goodman 2008). Peeking at intermediate results requires corrections (Johari et al. 2017).
6. Is performance differing significantly? If performance, or latency, is unexpectedly different, there may be caching issues or cold-start problems. A treatment may be slower due to caching issues of an LRU cache (Least Recently Used) and unbalanced variants (e.g., 90%/10%).
7. Are there experiment interactions? When an experimentation platform allows overlapping experiments, as most modern systems do, it is important to conduct a diagnostic to check all pairs of experiments for statistical interactions.
8. Is SUTVA violated? Experiments assume the Stable Unit Treatment Value Assumption (SUTVA) (Imbens and Rubin 2015), which states that experiment units (e.g., users) do not interfere with one another. Their behavior is impacted by their own variant assignment, and not by the assignment of others.

   The assumption could clearly be violated in settings, including the following:
   (a) Social networks, where a feature might spillover to a user's network.
   (b) Communication tools like Zoom, WebEx, and Skype involve peer-to-peer calls that often violate SUTVA.
   (c) Document authoring tools (such as, Microsoft Office and Google Docs) with co-authoring support.
   (d) Two-sided marketplaces (such as ad auctions, Airbnb, eBay, Lift, or Uber) can violate SUTVA through the "other" side. For example, lowering prices for Treatment has impact on Controls during auctions.
   (e) Shared resources (such as CPU, storage, and caches) can impact SUTVA (Kohavi and Longbotham 2010). If the Treatment leaks memory and causes processes to slow down due to garbage collection and possibly swapping of resources to disk, all variants suffer. If the Treatment crashes the machine in certain scenarios, those crashes also take down users who were in Control, so the delta on key metrics may not differ—both populations suffered similarly.
9. Are there heterogeneous treatment effects? The treatment effect may not be uniform and there may be segments that are impacted differently. For example, a JavaScript change may work on most browsers, but fail on an older version of Internet Explorer, leading to errors that may render the website unusable (Kohavi et al. 2020, Chapter 3: Twyman's Law and Experimentation Trustworthiness). A good overview of heterogenous treatment effects is available at EGAP (2018). Identifying interesting segments, or searching for interactions, can be done using machine learning and statistical techniques, such as Decision Trees (Athey and Imbens 2016) and Random Forests (Wager and Athey 2018).
10. Are there guardrail metrics to warn about unusual differences between variants? For example, changes to the fidelity of telemetry (e.g., click loss rate), cache hit rate differences, cookie write rates that differ, quick requests (Kohavi et al. 2020, Chapter 21; Zhao et al. 2016; Chen et al. 2019).
11. Is there a survivorship bias? Analyzing users who have been active for a long time (e.g., 2 months) introduces survivorship bias (Dmitriev et al. 2016).
12. Are robots (bots) introducing outliers or noise? Robots should be removed from any analysis of web data since their activity can

severely bias experiment results, see Kohavi et al. (2009b).

Experimentation Maturity Models

Experimentation maturity models (Fabijan et al. 2017, 2018; Optimizely 2018; Wider Funnel 2018; Brooks Bell 2015) consist of the phases organizations are likely to go through on the way to being data-driven and running every change through A/B experiments: Crawl, Walk, Run, and Fly.

As a rough rule of thumb, in the Crawl phase, an organization is running experiments approximately once a month (∼10/year), and it increases by 4–5x for each phase: organizations in the Walk phase will run experiments approximately once a week (∼50/year), Run is daily (∼250/year), and Fly is when you reach thousand(s)/year.

Experimentation Ethics

Any change could potentially be an unethical change. If so, it would also be unethical to experiment with that change. The Belmont Report (The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979) establishes principles for biomedical and behavioral studies, and to the Common Rule (Office for Human Research Protections 1991) establishes actionable review criteria based on these principles. One of the key points is "risk of substantial harm" and the concept of equipoise (Freedman 1987): whether the relevant expert community is in equipoise—genuine uncertainty—with respect to different treatments.

Any change that harms visitors or stakeholders in any way (physical, financial, emotional, psychological, social, or privacy) may be seen as unethical. In addition, if the content is dishonest, untruthful, misrepresenting or does not comply with the organization's explicit or implied agreement with users should be considered unethical.

Two litmus tests that we recommend are:

1. Could you ship the change to all users without a controlled experiment, given the organizational standards? If you could make the change to an algorithm, or to the look-and-feel of a product without an experiment, surely you should be able to run an experiment and scientifically evaluate the change first. Shipping code is, in fact, an experiment. It may not be a controlled experiment, but rather an inefficient sequential test where one looks at the time series; if key metrics (e.g., revenue, user feedback) are negative, the feature is rolled back.
2. If the experiment were published nationwide in a newspaper or a blog, would it be a public relations problem?

A couple of examples of experiments that were thought to be unethical by many people are the Facebook contagion experiment (Kramer et al. 2014) and the OKCupid experiment (Selterman 2014). Additional references on this topic are available (Kohavi et al. 2020; Loukides et al. 2018; FAT/ML 2019; ACM 2018; King et al. 2017; Benbunan-Fich 2017; Meyer 2018).

Summary

The internet and online connectivity of client software, websites, and online services provide a fertile ground for scientific testing methodology. Online experimentation is now recognized as a critical tool to determine whether a software or design change should be made. The benefit of experimenting online is the ability to set up a software platform for conducting the tests, which makes experimentation much more scalable and efficient and allows evaluating ideas quickly.

Cross-References

- A/B Tests
- Field Experiments
- Randomized Experiments
- Split Tests

Recommended Reading

ACM (2018) ACM code of ethics and professional conduct. June 22. https://www.acm.org/code-of-ethics

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. PNAS 113:7353–7360. https://doi.org/10.1073/pnas.1510489113

Benbunan-Fich R (2017) The ethics of online research with unsuspecting users: from A/B testing to C/D experimentation. Res Ethic 13(3–4):200–218. https://doi.org/10.1177/1747016116680664

Biau DJ, Jolles BM, Porcher R (2010) P value and the theory of hypothesis testing. Clin Orthop Relat Res 468(3):885–892

Bickel PJ, Doksum KA (1981) An analysis of transformations revisited. J Am Stat Assoc 76(374): 296–311. https://doi.org/10.1080/01621459.1981.10477649

Blank SG (2005) The four steps to the epiphany: successful strategies for products that win. Cafepress.com

Box GEP, Stuart Hunter J, Hunter WG (2005) Statistics for experimenters: design, innovation, and discovery, 2nd edn. Wiley

Bradley E, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York

Brooks Bell (2015) Click summit 2015 keynote presentation. Brooks Bell. http://www.brooksbell.com/wp-content/uploads/2015/05/BrooksBell_ClickSummit15_Keynote1.pdf

Casella G, Berger RL (2001) Statistical inference, 2nd edn. Cengage Learning

Chen N, Liu M, Ya X (2019) How A/B tests could go wrong: automatic diagnosis of invalid online experiments. In: WSDM '19 proceedings of the twelfth ACM international conference on web search and data mining. ACM, Melbourne, pp 501–509. https://dl.acm.org/citation.cfm?id=3291000

Crook T, Frasca B, Kohavi R, Longbotham R (2009) Seven pitfalls to avoid when running controlled experiments on the web. In: KDD '09 proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1105–1114

Deng A, Victor H (2015) Diluted treatment effect estimation for trigger analysis in online controlled experiments. WSDM 2015

Deng S, Longbotham R, Walker T, Ya X (2011) Choice of randomization unit in online controlled experiment. In: Joint statistical meetings proceedings, pp 4866–4877

Deng A, Xu Y, Kohavi R, Walker T (2013) Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. WSDM 2013

Deng A, Lu J, Litz J (2017) Trustworthy analysis of online A/B tests: pitfalls, challenges and solutions. In: WSDM the tenth international conference on web search and data mining. Cambridge

Deng A, Knoblich U, Jiannan L (2018) Applying the delta method in metric analytics: a practical guide with novel ideas. In: 24th ACM SIGKDD conference on knowledge discovery and data mining

Dmitriev P, Frasca B, Gupta S, Kohavi R, Vaz G (2016) Pitfalls of long-term online controlled experiments. In: IEEE international conference on big data. Washington, pp 1367–1376. https://doi.org/10.1109/BigData.2016.7840744

Dmitriev P, Gupta S, Kim DW, Vaz G (2017) A dirty dozen: twelve common metric interpretation pitfalls in online controlled experiments. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2017). ACM, Halifax, pp 1427–1436. https://doi.org/10.1145/3097983.3098024

Eckles D, Karrer B, Ugander J (2017) Design and analysis of experiments in networks: reducing bias from interference. J Causal Inference 5(1):23. https://www.deaneckles.com/misc/Eckles_Karrer_Ugander_Reducing_Bias_from_Interference.pdf

EGAP (2018) 10 things to know about heterogeneous treatment effects. EGAP: Evidence in Government and Politics. egap.org/methods-guides/10-things-heterogeneous-treatment-effects

Fabijan A, Dmitriev P, Olsson HH, Bosch J (2017) The evolution of continuous experimentation in software product development: from data to a data-driven organization at scale. In: ICSE '17 proceedings of the 39th international conference on software engineering. IEEE Press, Buenos Aires, pp 770–780. https://doi.org/10.1109/ICSE.2017.76

Fabijan A, Dmitriev P, McFarland C, Vermeer L, Olsson HH, Bosch J (2018) Experimentation growth: evolving trustworthy A/B testing capabilities in online software companies. J Softw: Evol Process 30(12):e2113. https://doi.org/10.1002/smr.2113

Fabijan A, Gupchup J, Gupta S, Omhover J, Qin W, Vermeer L, Dmitriev P (2019) Diagnosing sample ratio mismatch in online controlled experiments: a taxonomy and rules of thumb for practitioners. In: KDD '19 The 25th SIGKDD international conference on knowledge discovery and data mining. ACM, Anchorage

FAT/ML (2019) Fairness, accountability, and transparency in machine learning. http://www.fatml.org/

Fieller EC (1954) Some problems in interval estimation. J R Stat Soc Ser B 16(2):175–185. JSTOR 2984043 https://www.jstor.org/stable/2984043

Freedman B (1987) Equipoise and the ethics of clinical research. N Engl J Med 317(3):141–145. https://doi.org/10.1056/NEJM198707163170304

Georgiev G (2018) Analysis of 115 A/B tests: average lift is 4%, most lack statistical power analytics toolkit June 26. http://blog.analytics-toolkit.com/2018/analysis-of-115-a-b-tests-average-lift-statistical-power/

Georgiev, G (2019). Statistical methods in online A/B testing: Statistics for data-driven business decisions and risk management in e-commerce. Independently published. https://www.abtestingstats.com/

Good PI (2005) Permutation, parametric and bootstrap tests of hypotheses, 3rd edn. Springer

Goodman S (2008) A dirty dozen: twelve P-value misconceptions. Semin Hematol. https://doi.org/10.1053/j.seminhematol.2008.04.003

Goward C (2012) You should test that: conversion optimization for more leads, sales and profit or the art and science of optimized marketing. Sybex

Greenhalgh T (2014) How to read a paper: the basics of evidence-based medicine. BMJ Books. https://www.amazon.com/gp/product/B00IPG7GLC

Gupta S, Kohavi R, Tang D, Xu Y et al (2019) Top challenges from the first practical online controlled experiments summit. In: Dong XL, Teredesai A, Zafarani R (eds) SIGKDD explorations (ACM), vol 21 (1). https://bit.ly/OCESummit1

Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ (1995) Users' guides to the medical literature: IX. A method for grading health care recommendations. J Am Med Assoc 274(22):1800–1804. https://doi.org/10.1001/2Fjama.1995.03530220066035

Hochberg Y, Benjamini Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing Series B. J R Stat Soc 57(1): 289–300

Imbens GW, Rubin DB (2015) Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge University Press, Cambridge

Johari R, Pekelis L, Koomen P, Walsh D (2017) Peeking at A/B Tests. In: KDD '17 proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, Halifax, pp 1517–1525. https://doi.org/10.1145/3097983.3097992

Katzir L, Liberty E, Somekh O (2012) Framework and algorithms for network bucket testing. In: Proceedings of the 21st international conference on world wide web. pp 1029–1036

Kaushik A (2006) Experimentation and testing: a primer. Occam's Razor. http://www.kaushik.net/avinash/2006/05/experimentation-and-testing-a-primer.html. Accessed 22 May 2008

King R, Churchill EF, Tan C (2017) Designing with data: improving the user experience with A/B testing. O'Reilly Media

Kohavi R, Longbotham R (2010) Unexpected results in online controlled experiments. SIGKDD explorations. http://bit.ly/expUnexpected

Kohavi R, Thomke S (2017) The surprising power of online experiments. Harv Bus Rev 95:74–82. http://exp-platform.com/hbr-the-surprising-power-of-online-experiments/

Kohavi R, Longbotham R, Sommerfield D, Henne RM (2009a) Controlled experiments on the web: survey and practical guide. Data Min Knowl Disc 18: 140–181. http://bit.ly/expSurvey

Kohavi R, Crook T, Longbotham R (2009b) Online experimentation at microsoft. Third workshop on data mining case studies and practice prize. http://bit.ly/expMicrosoft

Kohavi R, Longbotham R, Walker T (2010) Online experiments: practical lessons. IEEE Computer, pp 82–85. http://bit.ly/expPracticalLessons

Kohavi R, Deng A, Frasca B, Longbotham R, Walker T, Ya X (2012) Trustworthy online controlled experiments: five puzzling outcomes explained. In: Proceedings of the 18th conference on knowledge discovery and data mining. http://bit.ly/expPuzzling

Kohavi R, Deng A, Frasca B, Walker T, Xu Y, Pohlmann N (2013) Online controlled experiments at large scale. In: KDD 2013 proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. http://bit.ly/ExPScale

Kohavi R, Deng A, Longbotham R, Ya X (2014) Seven rules of thumb for web site. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '14). http://bit.ly/expRulesOfThumb

Kohavi R, Tang D, Xu Y (2020) Trustworthy online controlled experiments: a practical guide to A/B testing. Cambridge University Press, Cambridge. https://experimentguide.com/

Kohavi R, Deng A, Vermeer L (2022) A/B testing intuition busters: Common misunderstandings in online controlled experiments. In: Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining (KDD '22), pp 3168–3177. https://doi.org/10.1145/3534678.3539160

Kramer ADI, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. Proc Natl Acad Sci 111:8788–8790. https://www.pnas.org/content/111/24/8788

Loukides M, Mason H, Patil DJ (2018) Ethics and data science. O'Reilly Media

Luca M, Bazerman MH (2020) The power of experiments: decision making in a data-driven world. The MIT Press. https://www.amazon.com/Power-Experiments-Decision-Making-Data-Driven/dp/0262043874

Malinas G, Bigelow J (2009) Simpson's paradox. Stanford encyclopedia of philosophy. http://plato.stanford.edu/entries/paradox-simpson/

Manzi J (2012) Uncontrolled: the surprising payoff of trial-and-error for business, politics, and society. Basic Books

Martin RC (2008) Clean code: a handbook of agile software craftsmanship. Prentice Hall

McFarland C (2012) Experiment!: website conversion rate optimization with A/B and multivariate testing. New Riders

McKinley D (2013) Testing to cull the living flower. http://mcfunley.com/testing-to-cull-the-living-flower

Meyer MN (2018) Ethical considerations when companies study – and fail to study – their customers. In: Selinger E, Polonetsky J, Tene O (eds) The Cambridge handbook of consumer privacy. Cambridge University Press, Cambridge

Moran M (2007) Do it wrong quickly: how the web changes the old marketing rules. IBM Press

Moran M (2008) Multivariate testing in action: quicken Loan's Regis Hadiaris on multivariate testing. December. www.biznology.com/2008/12/multivariate_testing_in_action/

Office for Human Research Protections (1991) Federal policy for the protection of human subjects ('Common Rule'). https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html

Optimizely (2018) Optimizely maturity model. https://www.optimizely.com/maturity-model/

Peterson ET (2004) Web analytics demystified: a Marketer's guide to understanding how your web site affects your business. Celilo Group Media and CafePress

Ries E (2011) The lean startup: how today's entrepreneurs use continuous innovation to create radically successful businesses. Crown Business

Rubin KS (2012) Essential scrum: a practical guide to the most popular agile process. Addison-Wesley Professional

Schrage M (2014) The innovator's hypothesis: how cheap experiments are worth more than good ideas. MIT Press

Selterman D (2014) The ethics of OKCupid's dating experiment. https://www.luvze.com/the-ethics-of-okcupids-dating-experiment/

Siroker D, Koomen P (2013) A/B testing: the most powerful way to turn clicks into customers. Wiley

Stone JV (2013) Bayes' rule: a tutorial introduction to Bayesian analysis. Sebtel Press

Tang D, Agarwal A, O'Brien D, Meyer M (2010) Overlapping experiment infrastructure: more, better, faster experimentation. In: KDD '10 the 16th ACM SIGKDD international conference on knowledge discovery and data mining

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979) The Belmont report April 18. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

Thomke SH (2020) Experimentation works: the surprising power of business experimentation. Harvard Business Review Press. https://www.amazon.com/Experimentation-Works-Surprising-Business-Experiments/dp/163369710X/

Ugander J, Karrer B, Backstrom L, Kleinberg J (2013) Graph cluster randomization: network exposure to multiple universes. In: KDD '13 proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, pp 329–337

Vickers AJ (2009) What is a p-value anyway? 34 stories to help you actually understand statistics. Pearson. https://www.amazon.com/p-value-Stories-Actually-Understand-Statistics/dp/0321629302

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 13(523):1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Wider Funnel (2018) The state of experimentation maturity 2018. Wider Funnel. https://www.widerfunnel.com/wp-content/uploads/2018/04/State-of-Experimentation-2018-Original-Research-Report.pdf

Xie H, Aurisset J (2016) Improving the sensitivity of online controlled experiments: case studies at Netflx. In: KDD '16 proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. Association for Computing Machinery, pp 645–654

Xu Y, Chen N, Fernandez A, Sinno O, Bhasin A (2015) From infrastructure to culture: A/B testing challenges in large scale social networks. In: KDD '15 proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, Sydney, pp 2227–2236. https://doi.org/10.1145/2783258.2788602

Zhao Z, Chen M, Matheson D, Stone M (2016) Online experimentation diagnosis and troubleshooting beyond AA validation. In: DSAA 2016 IEEE international conference on data science and advanced analytics. IEEE, pp 498–507. https://ieeexplore.ieee.org/document/7796936